What is Special About Natural Language Generation Research?

William C. Mann
USC Information Sciences Institute[1]
Marina del Rey, CA

Since the guidance given to the panel was more provocative than regulative, I have organized my statement around just one of the questions:

What is special about Text Generation relative to NL Understanding?

This breaks down conveniently into parts:

1. Are there foundational ideas that generation and understanding work share?

2. What are the technical distinctives of generation?

3. Are those distinctives real?

4. What are the special characteristics of generation as a research task?

1 Shared Foundations

While there are considerable differences in the tasks to be solved in Text Generation and NL Understanding, the two areas of research draw on a significant number of shared ideas and knowledge.[2] They constitute an account of what the facts and phenomena of natural language are. Moving from fine-grained to coarse-grained phenomena, they include:

1. Lexicon: Most work in both understanding and generation assumes a taxonomy of basic word classes, a notion of the semantic senses of words and a morphology. Also in both, there is currently a strong trend toward recognizing many sorts of lexical complexities: idioms, collocations, lexical functions (in several senses) and other inter-item interactions.

2. Grammar: Ther are shared descriptions of the types of constructions that are available in a specific language. At a minimum, a language processing

[1]Legal Notice: This research was supported by the Air Force Office of Scientific Research contract No. F49620-84-C-0100. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research of the U.S. Government.

[2]Bob Kasper contributed heavily to this section.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1987		**00-00-1987 to 00-00-1987**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
What is Special About Natural Language Generation Research?	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Information Sciences Institute,University of Southern California,4676 Admirality Way Suite 1001,Marina del Rey,CA,90292**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**5**	

program includes a grammar, some specification of a set of syntactic patterns.

3. Discourse Phenomena: Descriptions of various discourse phenomena are important in both lines of work. Anaphora is particularly prominent. A cluster of phenomena identified with terms such as theme, focus and topic is also basic. There is also a general recognition that ordinary language does not make explicit everything that is being conveyed, and that the non-explicit material is just as important as the explicit material in effective language use. It seems likely that there will be substantial cross-fertilization from these two lines of current work on discourse, partly because the available descriptions of discourse are still not well agreed upon.

4. Situational Phenomena: The situation in which the language is used, including a description of the language user and the task at hand, are acknowledged as important and actively studied. Goal pursuit by the language user(s) is regarded as an important orienting notion.

Both generation and understanding are working hard on all of these. Inevitably, there is some complemenarity (see Section 3.) But although the descriptive foundations are shared in a loose way, we will see that the sorts of problems addressed differ sharply.

More substantial sharing occurs in the areas of knowledge representation and inference. Here the problems and solutions, not just the recognition of phenomena, are shared. There is hope for convergence, for one all-sufficient underlying representational form, and for a non-directional view of language. It is often suggested that an adequate text generator must have an understander inside to check its work. Still, the research activity is dominated by the differences rather than the shared elements.

2 Technical Distinctives of Text Generation

Just observing work on understanding and generation, it's clear that the people working and writing on these topics are usually not writing about the same things. To start to understand the situation we can look at the technical differences and then later judge how fundamental these differences are.

What are the apparent differences? One class consists of problems which are major sources of difficulty in NL Understanding but which are minor or absent in NL Generation:

1. Covering all the ways to say things is not a problem. These days it's sufficient (and difficult enough) to have one way to say everything, with just enough perturbations to get sufficiently fluent text.

2. Goal identification is not a problem. A generation system can know its own goals easily. Of course, coming up with the right goals is still a problem.

3. Vocabulary coverage is not a problem. The lexicon of a generator can be created in correspondence with available knowledge; the user's unbounded number of other ways of expressing the knowledge do not have an impact.

4. Ambiguity is a secondary problem. People, operating in context with a rich knowledge of the subject matter, can disambiguate generated language very well.

Another class consists of problems which are important in Generation but minor or absent in Understanding:

1. Deciding how much to say, and what things to not say, are problems. This involves maintaining brevity, avoiding saying what is too obvious, and yet providing sufficient background information to make the generated text comprehensible.

2. Design of text structure is a problem. This is sometimes taken to be the coherence problem as well: text must be coherent, and appropriate structure makes it so. Structure design has many identifiable subproblems:

    a. Structure building includes adding material to make presentation of the basic subject matter work. For example, it is often necessary to add evidence, concessives, circumstantials, antithesis, contrast and other supporting material.

    b. Structuring a text causes assertion-like effects in addition to the expected effects of individual clauses. Controlling these effects, and taking advantage of them as a resource, is a problem.

    c. Ordering the material for presentation is very consequential.

    d. Various sorts of text carry the expectation of special patterns and formulaic text: titles, abstracts, salutations, origination dates, authorship notes and acknowledgments.

    e. Making the text smooth flowing and easy to comprehend involves leading the reader's attention. There are many particular techniques which contribute. This requirement constrains structure design and requires extra work at the structural and sentential levels.

3. Even after creating a detailed text plan, with all clauses identified, there are substantial additional technical issues in carrying out the plan.

    a. Presuming that the plan is in terms of a sequence of (effects of)

clauses, the sentence boundaries are not determined. Which clauses should be combined into sentences? What relations need to be expressed by conjunctions? What conjunction uses can be reduced to noun conjunction or some other lower rank?

b. Deciding when to use anaphora is a problem.

c. Lexical selection is a problem. Related, there are many varieties of idioms and lexical colocations whose restricted character is important only for generation, not understanding.

d. English has rather elaborate provisions which enable the reader's attention to flow smoothly over the material. These include emphasis devices, and also various kinds of theme control (including passivization as one of many kinds). These must be controlled in order to create high quality running text.

3 The Alternative View: The Differences in the Tasks are Unreal

The claim has been made that there are really no underlying language problems that are unique to either generation or understanding. Rather, every evident problem has a counterpart which may or may not be evident on the other side of the fence. So, for example, the counterpart of (Generation: deciding how much to say) is (Understanding: identify the selectivity involved in saying just this much.) The counterpart of (Generation: lexical selection) is (Understanding: drawing conclusions from the fact that this particular term was used rather than alternative terms.) And so forth. The underlying claim is that if a process is used in generation, it has effects which may be discernible, interpretable, even significant. The earliest use of this claim that I know was by Chip Bruce, in the presentation of [Bruce 75].

As a statement of what sorts of effects can (in principle) be found, this has a certain plausibility, and may be technically correct. Nevertheless, it does not represent the state of the art in terms of problems actually worked on. Instead, the lists of problems being addressed by generation and understanding research differ substantially, and will remain different for a long time to come. This is because the problems that limit the achievable quality of performance, the problems that pace progress, differ strongly between generation and understanding.

4 Distinctives of Text Generation as a Research Task

There are non-technical factors that make research into text generation very different from text understanding research:

1. In both duration and number of workers, there has been far less activity in generation than in understanding. In spite of much recent expansion in

generation work (see [Kempen 86] for a representative collection) there are far fewer precedents and established results in generation. Work in generation is less known, so much so that some people habitually conceive of all AI language research as NLU (natural language understanding.) (See, for example, the IJCAI87 call for papers.)

2. It is easier to control a generation task, since it is not subject to an uncontrolled input source (the user.) There is inherently more control over vocabulary, lexical phenomena, syntactic range and semantic diversity in generation.

3. Generation and understanding need to be understood in terms of an overall model of human communication. The nature of language and the constraints on its use come from its role in communication. If investigation of communication is taken as the underlying task, then generation gives much better access to that task, just because it is much easier to develop methods and programs that work with whole discourses rather than being restricted to tiny numbers of sentences.

5 Conclusions

The set of technical problems that limit the quality of generated text is very different from the corresponding set of problems that limits the quality of natural language understanding. While the problems might, in principle, be problems of both understanding and generation, they are not so in practise.

Generation provides some important advantages over understanding as a research subject, because it does not require coping with an uncontrolled language-user. As a result, research into computational models of communication can sometimes be made more efficient by studying generation.

References

[Bruce 75] Bruce, B. C., "Generation as a social action," in *Proceedings of Theoretical Issues in Natural Language Processing-I (TINLAP)*, pp. 64-67, Cambridge, Mass., June 1975.

[Kempen 86] Kempen, Gerard, (ed.), *Proceedings of the Third International Workshop on Text Generation*, , Nijmegen, The Netherlands, 1986.